

A multi-dimensional stream and its signature representation

Hao Ni

September 14, 2015

Abstract

The signature of a path is an essential object in the theory of rough paths. The signature representation of the data stream can recover standard statistics, e.g. the moments of the data stream. The classification of random walks indicates the advantages of using the signature of a stream as the feature set for machine learning.

1 Introduction

This short paper is devoted to show that the signature of the lead-lag transformation is a useful way to encode a multi-dimensional unstructured data stream. We aim to demonstrate the following points:

1. The signature of a discrete sample stream is a rich statistics and encodes the essential information of data stream;
2. The truncated signature of a discrete sample stream provides a summary in terms of the effect of this stream and it leads to dimension reduction for this original stream;
3. The signature of a discrete sample can be used for parameter inference and prediction.

The main result is Theorem 4.1, which states that no matter how frequently the path is sampled, the p^{th} moment of the increment process is a linear functional on the truncated signature up to degree p .

2 Notation and Preliminaries

2.1 Signatures

Let us start with introducing the tensor algebra space, in which the signature of a path takes value.

Definition 2.1 (Tensor algebra space) *A formal E -tensor series is a sequence of tensors $(a_n \in E^{\otimes n})_{n \in \mathbb{N}}$ which we write $a = (a_0, a_1, \dots)$. There are two binary operations on E -tensor series, an addition $+$ and a product \otimes , which are defined as follows. Let $\mathbf{a} = (a_0, a_1, \dots)$ and $\mathbf{b} = (b_0, b_1, \dots)$ be two E -tensor series. Then we define*

$$\mathbf{a} + \mathbf{b} = (a_0 + b_0, a_1 + b_1, \dots), \quad (1)$$

and

$$\mathbf{a} \otimes \mathbf{b} = (c_0, c_1, \dots), \quad (2)$$

where for each $n \geq 0$,

$$c_n = \sum_{k=0}^n a_k \otimes b_{n-k}. \quad (3)$$

The product $\mathbf{a} \otimes \mathbf{b}$ is also denoted by \mathbf{ab} . We use the notation $\mathbf{1}$ for the series $(1, 0, \dots)$, and $\mathbf{0}$ for the series $(0, 0, \dots)$. If $\lambda \in \mathbb{R}$, then we define $\lambda \mathbf{a}$ to be $(\lambda a_0, \lambda a_1, \dots)$.

Definition 2.2 The space $T((E))$ is defined to be the vector space of all formal E -tensors series.

Similar to the real valued case, we can define the exp mapping on $T((E))$ as follows.

Definition 2.3 Let \mathbf{a} be arbitrary element of $T((E))$. Then $\exp(\mathbf{a})$ is the element of $T((E))$ by

$$\exp(\mathbf{a}) := \sum_{n=0}^{\infty} \frac{\mathbf{a}^{\otimes n}}{n!}.$$

Now we are in a position to give the definition of the signature of a path of bounded variation (finite length).

Definition 2.4 (Signature of a path) Let J be a compact interval and X be a continuous function of finite length, which maps J to E . The signature $S(X)$ of X over the time interval J is an element $(1, X^1, \dots, X^n, \dots)$ of $T((E))$ defined for each $n \geq 1$ as follows

$$X^n = \int \cdots \int_{\substack{u_1 < \dots < u_n, \\ u_1, \dots, u_n \in J}} dX_{u_1} \otimes \dots \otimes dX_{u_n},$$

where the integration is in the sense of Young's integral. The truncated signature of X of order n is denoted by $S^n(X)$, i.e. $S^n(X) = (1, X^1, \dots, X^n)$, for every $n \in \mathbb{N}$.

Remark 2.1 Suppose that $\{e_i\}_{i=1}^d$ be a basis of E , and thus for every $n \geq 0$, $\{e_{i_1} \otimes \dots \otimes e_{i_n}\}_{i_1, \dots, i_n \in \{1, \dots, d\}}$ forms a basis of $E^{\otimes n}$. Therefore $S(X)$ can be rewritten as follows:

$$S(X) = 1 + \sum_{n=1}^{\infty} \sum_{\substack{i_1, \dots, i_n \\ \in \{1, \dots, d\}}} \left(\int \cdots \int_{\substack{u_1 < \dots < u_n \\ u_1, \dots, u_n \in J}} dX_{u_1}^{(i_1)} dX_{u_2}^{(i_2)} \dots dX_{u_n}^{(i_n)} \right) e_{i_1} \otimes e_{i_2} \otimes \dots \otimes e_{i_n}.$$

The signature of a path can be simply regarded as a formal infinite sum of non-commutative tensor products, and the coefficient of each monomial is determined by its corresponding coordinate iterated integral. For every multi-index $I = (i_1, \dots, i_n)$, denote by \mathbf{X}^I the following iterated integral of X indexed by I , i.e.

$$\mathbf{X}^I = \int \cdots \int_{\substack{u_1 < \dots < u_n \\ u_1, \dots, u_n \in J}} dX_{u_1}^{(i_1)} dX_{u_2}^{(i_2)} \dots dX_{u_n}^{(i_n)}.$$

The first property is Chen's identity (Theorem 2.1), which asserts that the signature of the concatenation of two paths is the tensor product of the signature of each path.

Definition 2.5 Let $X : [0, s] \rightarrow E$ and $Y : [s, t] \rightarrow E$ be two continuous paths. Their concatenation is the path $X * Y$ defined by

$$(X * Y)_u = \begin{cases} X_u, & u \in [0, s]; \\ X_s + Y_u - Y_s, & u \in [s, t], \end{cases}$$

where $0 \leq s \leq t$.

Theorem 2.1 (Chen's identity) Let $X : [0, s] \rightarrow E$ and $Y : [s, t] \rightarrow E$ be two continuous paths with finite 1-variation. Then

$$S(X * Y) = S(X) \otimes S(Y), \quad (4)$$

where $0 \leq s \leq t$.

The proof can be found in [2].

Let $\{e_i^*\}_{i=1}^d$ be a basis of the dual space E^* . Then for every $n \in \mathbb{N}$, $\{e_{i_1}^* \otimes \cdots \otimes e_{i_n}^*\}$ it can be naturally extended to $(E^*)^{\otimes n}$ by identifying the basis $(e_I = e_{i_1}^* \otimes \cdots \otimes e_{i_n}^*)$ as

$$\langle e_{i_1}^* \otimes \cdots \otimes e_{i_n}^*, e_{j_1} \otimes \cdots \otimes e_{j_n} \rangle = \delta_{i_1, j_1} \cdots \delta_{i_n, j_n}.$$

The linear action of $(E^*)^{\otimes n}$ on $E^{\otimes n}$ extends naturally to a linear mapping $(E^*)^{\otimes n} \rightarrow T((E))^*$ defined by

$$e_I(\mathbf{a}) = e_I^*(a_n),$$

where $I = (i_1, \dots, i_n)$.

Hence the linear forms e_I^* , as I span the set of finite words in the letters $1, \dots, d$ form a basis of $T(E^*)$. Let $T((E))^*$ denote the space of linear forms on $T((E))$ induced by $T(E^*)$. Let us consider a word $I = (i_1, \dots, i_n)$, where $i_1, \dots, i_n \in \{1, \dots, d\}$. Define π^I as e_I^* restricting the domain to the range of the signatures, denoted by $S(\mathcal{V}^1[0, T], E)$, in formula

$$\pi^I(S(X)) = e_I^*(S(X)),$$

where X is any E -valued continuous path of bounded variation.

For any two words I and J , the pointwise product of two linear forms π^I and π^J as real valued functions is a quadratic form on $S(\mathcal{V}^1[0, T], E)$, but it is remarkable that it is still a linear form, which is stated in Theorem 2.2. Let us introduce the definition of the shuffle product.

Definition 2.6 We define the set $S_{m,n}$ of (m, n) shuffles to be the subset of permutation in the symmetric group S_{m+n} defined by

$$S_{m,n} = \{\sigma \in S_{m+n} : \sigma(1) < \cdots < \sigma(m), \sigma(m+1) < \cdots < \sigma(m+n)\}.$$

Definition 2.7 The shuffle product of π^I and π^J denoted by $\pi^I \sqcup \pi^J$ defined as follows:

$$\pi^I \sqcup \pi^J = \sum_{\sigma \in S_{m,n}} \pi^{(k_{\sigma^{-1}(1)}, \dots, k_{\sigma^{-1}(m+n)})},$$

where $I = (i_1, i_2, \dots, i_n)$, $J = (j_1, j_2, \dots, j_m)$ and $(k_1, \dots, k_{m+n}) = (i_1, \dots, i_n, j_1, \dots, j_m)$.

Theorem 2.2 (Shuffle Product Property) Let X be a path of bounded variation. Let I and J be two arbitrary indices. The following identity holds:

$$\pi^I(S(X))\pi^J(S(X)) = (\pi^I \sqcup \pi^J)(S(X)).$$

3 A discrete sampled path and the signature of its lead-lag transformation

In the following we constrain our discussion on paths observed at a finite number of time stamps and take value in $E := \mathbb{R}^d$.

3.1 The discrete sampled path and the lead-lag transformation

Let $\{x_n\}_{n=1}^L$ be an increment process, where $x_n \in E$. (You can think of it as a return process.) Let $\mathbf{X} := \{X_n\}_{n=0}^L$ denote the corresponding partial sum process of $\{x_n\}_{n=0}^{L-1}$. (It can be thought as a price process.) Mathematically, \mathbf{X} is defined as follows:

$$\begin{aligned} X_0 &= 0; \\ X_{n+1} &= \sum_{i=1}^n x_i, \text{ if } n = 1, \dots, L. \end{aligned}$$

Now let us introduce the lead-lag transformation associated with a d -dimensional stream \mathbb{X} ([1]).

Definition 3.1 (Lead-Lag Transformation) Let $\mathbb{X} := \{X_n\}_{n=0}^L$ be a d -dimensional discrete sampled path. The lead-lag transformation associated with \mathbf{X} is a $2d$ -dimensional path which is obtained by linear interpolation of $\mathbf{X} := \{X_n\}_{n=0}^L$, where $\mathbf{X}_0^{(i)} = X_0^{(i)}$ and $\mathbf{X}_{2n-1}^{(i)} = X_n^{(i)}$ and for every $n \in \{0, \dots, L-1\}$ and for every $i \in \{1, \dots, d\}$,

$$\begin{aligned} \mathbb{X}_{2n+2}^{(i)} &= \mathbb{X}_{2n+1}^{(i)} = X_{n+1}^{(i)} \\ \mathbb{X}_{2n}^{(i+d)} &= \mathbb{X}_{2n+1}^{(i+d)} = X_n^{(i)}. \end{aligned}$$

Let \mathcal{L} denote the lead-lag transformation operator.

The lead-lag process \mathbb{X} is in the form of the following:

$$\begin{pmatrix} \mathbb{X}_0, \\ \parallel \\ X_0^{(1)} \\ X_0^{(2)} \\ \vdots \\ X_0^{(d)} \\ X_0^{(1)} \\ X_0^{(2)} \\ \vdots \\ X_0^{(d)} \end{pmatrix}, \quad \begin{pmatrix} \mathbb{X}_1, \\ \parallel \\ X_1^{(1)} \\ X_1^{(2)} \\ \vdots \\ X_1^{(d)} \\ X_0^{(1)} \\ X_0^{(2)} \\ \vdots \\ X_0^{(d)} \end{pmatrix}, \quad \begin{pmatrix} \mathbb{X}_2, \\ \parallel \\ X_1^{(1)} \\ X_1^{(2)} \\ \vdots \\ X_1^{(d)} \\ X_1^{(1)} \\ X_1^{(2)} \\ \vdots \\ X_1^{(d)} \end{pmatrix} \quad \cdots \quad \begin{pmatrix} \mathbb{X}_{2n-1}, \\ \parallel \\ X_n^{(1)} \\ X_n^{(2)} \\ \vdots \\ X_n^{(d)} \\ X_{n-1}^{(1)} \\ X_{n-1}^{(2)} \\ \vdots \\ X_{n-1}^{(d)} \end{pmatrix}, \quad \begin{pmatrix} \mathbb{X}_{2n}, \\ \parallel \\ X_n^{(1)} \\ X_n^{(2)} \\ \vdots \\ X_n^{(d)} \\ X_n^{(1)} \\ X_n^{(2)} \\ \vdots \\ X_n^{(d)} \end{pmatrix}$$

Lemma 3.1 (The multiplicative of the lead-lag transformation) For any two discrete sampled path $\mathbf{X} = \{X_n\}_{n=0}^{L_1}$ and $\mathbf{Y} = \{Y_n\}_{n=0}^{L_2}$

$$\mathcal{L}(\mathbf{X} * \mathbf{Y}) = \mathcal{L}(\mathbf{X}) * \mathcal{L}(\mathbf{Y}),$$

where $\mathbf{X} * \mathbf{Y}$ denote the concatenation of two discrete sampled path, i.e.

$$(\mathbf{X} * \mathbf{Y})_n = \begin{cases} X_n & \text{if } n \leq L_1 - 1 \\ X_{L_1} - Y_0 + Y_{n-L_1} & \text{if } L_1 \leq n \leq L_1 + L_2. \end{cases}$$

3.2 The signature of the lead-lag transformation

Let us define the signature of the discrete sampled stream, and discuss the relevant properties.

Definition 3.2 (The signature representation of a discrete sampled stream)

Let \mathbf{X} be a discrete sampled path in E and \mathbb{X} is the lead-lag transformation of \mathbf{X} . The signature of \mathbf{X} is defined to be the signature of \mathbb{X} , denoted by $S(\mathbb{X})$. Let $S_d(\mathbb{X})$ denote the truncated signature of \mathbb{X} up to degree d . Let \mathcal{DS} denote the range of signatures of the lead-lag transformation of discrete sampled paths in E .

Lemma 3.2 (Chen's Identity for Discrete Sampled Path) For any two discrete sampled path $\mathbf{X} = \{X_n\}_{n=0}^{L_1}$ and $\mathbf{Y} = \{Y_n\}_{n=0}^{L_2}$.

$$S(\mathcal{L}(\mathbf{X} * \mathbf{Y})) = S(\mathcal{L}(\mathbf{X})) \otimes S(\mathcal{L}(\mathbf{Y})).$$

Definition 3.3 (Additive functional on \mathcal{DS}) Let K be a linear form on $T((E))$. We say that K is additive in \mathcal{DS} if and only if for every $S(\mathbb{X}), S(\mathbb{Y}) \in \mathcal{DS}$, it follows that

$$K(S(\mathbb{X} * \mathbb{Y})) = K(S(\mathbb{X})) + K(S(\mathbb{Y})).$$

For convenience, let us adopt the following notation

Definition 3.4 Fix any positive integer p . Let $\mathcal{K}_I^{(p)}$ denote the set of the linear forms on $T((E))$ such that it can be written as

$$\sum_{|J|=p, J=(J_1, I)} C_J \pi^{(J)}$$

where C_J are all constants and the summation is taken over all J such that J is of length p and ended in the substring I .

4 One Dimensional Stream Case

Let us focus on one dimensional case, and we will show that the signature of \mathbb{X} contains rich information of the path \mathbf{X} and it is a good basis function to represent the standard statistic, for example, the empirical moments of increments of \mathbf{X} (Theorem 4.1). Let us start with discussion on properties of the signature of \mathbb{X} .

By Chen's identity and simple calculation, the signature of a path in \mathcal{DS} can be given so explicit as follows:

Lemma 4.1 (Signature of one-dimensional discrete path) *For any $\mathbb{X} \in \mathcal{DS}$, and $\{x_i\}_{i=1}^L$ is the increment process associated with \mathbb{X} , then*

$$S(\mathbb{X}) = \bigotimes_{i=1}^L \exp(x_i e_1) \otimes \exp(x_i e_2)$$

Lemma 4.2 *For every index I ending in 2 and any positive integer p , there exists $K \in \mathcal{K}_2^{(|I|+p)}$, for any $\mathbb{X}_L \in \mathcal{DS}$, such that*

$$\pi^{(I, M_p)}(S(\mathbb{X}_L)) = K(S(\mathbb{X}_L)).$$

where M_p is p copies of 1.

For every index I ending in 1 and any positive integer p , there exists $K \in \mathcal{K}_1^{(|I|+p)}$, for any $\mathbb{X}_L \in \mathcal{DS}$, such that

$$\pi^{(I, K_p)}(S(\mathbb{X}_L)) = K(S(\mathbb{X}_L)).$$

where $M_p = (1, \dots, 1)$, i.e. p copies of 1.

Proof. First of all, let us prove that the case $p = 1$. As I ends in 2, then we can rewrite I as $(J, 2)$. Since $(\pi^{(1)} - \pi^{(2)})(S(\mathbb{X})) = 0$, then

$$\begin{aligned} 0 &= \pi^I(\pi^{(1)} - \pi^{(2)}) = \pi^{(J \sqcup 1, 2)} + \pi^{(I_2, 1)} - \pi^{I_2} \sqcup \pi^{(2)}. \\ \pi^{(I, 1)} &= \pi^I \sqcup \pi^{(2)} - \pi^{(J \sqcup 1, 2)} \in \mathcal{K}_2^{|I|+p}. \end{aligned}$$

Then we prove this statement by induction on p . Let K_p be p copies of 2s.

$$0 = \pi^I(\pi^{(M_p)} - \pi^{(K_p)}) = \pi^{(J \sqcup M_p, 2)} + \pi^{(I \sqcup M_{p-1}, 1)} - \pi^I \sqcup \pi^{K_p}.$$

Let us investigate the term $\pi^{(I \sqcup M_{p-1}, 1)}$.

$$(I \sqcup M_{p-1}, 1) = (I, M_p) + \sum_{k=1}^{p-1} (J \sqcup M_k, 2, M_{p-k}),$$

and thus

$$\pi^{(I \sqcup M_{p-1}, 1)} = \pi^{(I, M_p)} + \sum_{k=1}^{p-1} \pi^{(J \sqcup M_k, 2, M_{p-k})}.$$

For any $k = 1, \dots, p-1$, by induction hypothesis, there exist the linear functional $G \in \mathcal{K}_2^{|I|+p}$ such that for any $S(\mathbb{X}) \in \mathcal{DS}$,

$$\pi^{(J \sqcup M_k, 2, M_{p-k})} S(\mathbb{X}) = G(S(\mathbb{X})).$$

Therefore

$$\begin{aligned} \pi^{(I, M_p)} &= \pi^I \sqcup \pi^{K_p} - \pi^{(J \sqcup M_p, 2)} - \sum_{k=1}^{p-1} \pi^{(J \sqcup M_k, 2, M_{p-k})}, \\ &= \pi^I \sqcup \pi^{K_p} - \pi^{(J \sqcup M_p, 2)} - G \in \mathcal{K}_2^{|I|+p}. \end{aligned}$$

Now we complete the first part of the statement. We can use the same strategy to show the second part of the statement. ■

Remark 4.1 Since $\pi^{M_p} = \pi^{K_p}$, Lemma 4.2 shows that for each index I , $\pi^{(I)}$ can be rewritten as a linear functional in $\mathcal{K}_2^{|I|}$.

Lemma 4.3 For any index $I = (i_1, \dots, i_{n-1}, 2)$, and any $S(\mathbb{X}_L) \in \mathcal{DS}$,

$$\pi^{(I, 1)}(S(\mathbb{X}_L)) = \sum_{j=1}^L \pi^I(S(\mathbb{X}_{j-1}))x_j. \quad (5)$$

Proof. We show this lemma by induction on L . For $L = 1$, both sides of 5 are equal to 0. By Chen's identity, for $L \geq 1$, it follows that

$$\begin{aligned} \pi^{(I, 1)}(S(\mathbb{X}_L)) &= \pi^{(I, 1)}(S(\mathbb{X}_{L-1}) \otimes S(\mathbb{X}_{L-1, L})) \\ &= \pi^{(I, 1)}(S(\mathbb{X}_{L-1})) + \pi^{(I)}(S(\mathbb{X}_{L-1}))x_L \end{aligned}$$

because

$$S(\mathbb{X}_{L-1, L}) = \exp(x_L e_1) \otimes \exp(x_L e_2)$$

Then it follows by the induction hypothesis that

$$\begin{aligned} \pi^{(I, 1)}(S(\mathbb{X}_L)) &= \sum_{j=1}^{L-1} \pi^I(S(\mathbb{X}_{j-1}))x_j + \pi^{(I)}(S(\mathbb{X}_{L-1}))x_L \\ &= \sum_{j=1}^L \pi^I(S(\mathbb{X}_{j-1}))x_j. \end{aligned}$$

■

Lemma 4.4 For any index $I = (i_1, \dots, i_{n-1}, 2)$ and $k \geq 1$ there exists a linear functional F depending only on I and k , and $F \in \mathcal{K}_2^{n+k}$ such that for any $S(\mathbb{X}_L) \in \mathcal{DS}$, it holds that

$$F(S(\mathbb{X}_L)) = \sum_{j=1}^L \pi^I(S(\mathbb{X}_{j-1}))x_j^k. \quad (6)$$

Proof. For $k = 1$, it is proved in Lemma 4.3. Assume that $k \leq K - 1$ is true. Let us consider the case where $k = K$.

$$\begin{aligned} & \pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_L)) - \pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_{L-1})) \\ &= \sum_{j=1}^k \pi^{(I_2, 1^{*j})}(S(\mathbb{X}_{L-1})) \frac{x_L^{k-j}}{(k-j)!}. \end{aligned}$$

After rearranging the above formula we have that

$$\pi^{(I_2)}(S(\mathbb{X}_{L-1}))x_L^k = k! \left(\pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_L)) - \pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_{L-1})) + \sum_{j=1}^{k-1} \pi^{(I_2, 1^{*j})}(S(\mathbb{X}_{L-1})) \frac{x_L^{k-j}}{(k-j)!} \right).$$

By telescope sum of the above equation, we have that

$$\begin{aligned} & \sum_{i=1}^L \pi^{(I_2)}(S(\mathbb{X}_{i-1}))x_i^k \\ &= k! \pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_L)) + k! \sum_{i=1}^L \left(\sum_{j=1}^{k-1} \pi^{(I_2, 1^{*j})}(S(\mathbb{X}_{i-1})) \frac{x_i^{k-j}}{(k-j)!} \right) \\ &= k! \pi^{(I_2, 1, \dots, 1)}(S(\mathbb{X}_L)) + k! \left(\sum_{j=1}^{k-1} \frac{1}{(k-j)!} \sum_{i=1}^L \pi^{(I_2, 1^{*j})}(S(\mathbb{X}_{i-1}))x_i^{k-j} \right) \end{aligned}$$

By Lemma 4.2, there is a linear functional G depending on $(I_2, 1^{*j})$ and $k - j$, such that

$$\pi^{(I_2, 1^{*j})} = G.$$

Then by induction hypothesis,

$$\sum_{i=1}^L \pi^{(I_2, 1^{*j})}(S(\mathbb{X}_{i-1}))x_i^{k-j}$$

can be rewritten as a linear function on \mathcal{K}_2^{n+k} . $\pi^{(I_2, 1, \dots, 1)}$ can be rewritten as a linear functional in \mathcal{K}_2^{n+k} , so is $\sum_{i=1}^L \pi^{(I_2)}(S(\mathbb{X}_{i-1}))x_i^k$. Now the proof is complete. \blacksquare

Lemma 4.5 Let $L_1 \in \mathcal{K}_1^{(p)}$ and L_1 is additive, then there exists $\tilde{L}_1 \in \mathcal{K}_1^{(p+2)}$, such that

$$\tilde{L}_1(S(\mathbb{X}_n)) = - \sum_{i=1}^n L_1(S(\mathbb{X})_i) \frac{x_i^2}{2};$$

Proof. Let $L_1 := \sum_{I_1} C_{I_1} \pi^{(I_1)}$. For $n \geq 1$, it holds that

$$\begin{aligned} & \pi^{(I_1, 2, 1)}(S(\mathbb{X}_n)) \\ &= \pi^{(I_1, 2, 1)}(S(\mathbb{X}_{n-1}) \otimes S(\mathbb{X}_{n-1, n})) \\ &= \pi^{(I_1, 2, 1)}(S(\mathbb{X}_{n-1})) + \pi^{(I_1, 2)}(S(\mathbb{X}_{n-1}))x_n. \end{aligned}$$

Similarly we have

$$\begin{aligned}
& \pi^{(I_1, 2, 2)}(S(\mathbb{X}_n)) = \pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1}) \otimes S(\mathbb{X}_{n-1, n})) \\
&= \pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1})) + \pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1, n})) \\
&\quad + \pi^{(I_1, 2)}(S(\mathbb{X}_{n-1}))\pi^{(2)}(S(\mathbb{X}_{n-1, n})) + \pi^{(I_1)}(S(\mathbb{X}_{n-1}))\pi^{(2, 2)}(S(\mathbb{X}_{n-1, n})) + \mathcal{R}_{I_1(S(\mathbb{X}_{n-1}), x_n)} \\
&= \pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1})) + \pi^{(I_1)}(S(\mathbb{X}_{n-1, n}))\frac{x_n^2}{2} \\
&\quad + \pi^{(I_1)}(S(\mathbb{X}_{n-1}))\frac{x_n^2}{2} + \pi^{(I_1, 2)}(S(\mathbb{X}_{n-1}))x_n \\
&\quad + \mathcal{R}_{I_1}(S(\mathbb{X}_{n-1}), x_n).
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{R}_{I_1}(S(\mathbb{X}_{n-1}), x_n) &= \sum_{J * J_1 = I_1, J \neq \emptyset, J_1 \neq \emptyset} \pi^{(J)}S(\mathbb{X}_{n-1})\pi^{J_1}(S(\mathbb{X}_{n-1, n})) \\
&= \sum_{J * J_1 = I_1, J \neq \emptyset, J_1 \neq \emptyset} \pi^{(J)}S(\mathbb{X}_{n-1})c_{J_1}x_n^{|J_1|}.
\end{aligned}$$

The last equality comes from the fact that

$$\begin{aligned}
\pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1, n})) &= \pi^{(I_1, 2, 2)}(\exp(x_n e_1) \otimes \exp(x_n e_2)) \\
&= \pi^{(I_1)}(\exp(x_n e_1))\pi^{(2, 2)}(\exp(x_n e_2)) \\
&= \pi^{(I_1)}(S(\mathbb{X}_{n-1, n}))\frac{x_n^2}{2}.
\end{aligned}$$

By Lemma 4.4, there exists a linear functional G_{I_1} on \mathcal{K}_1^{p+2} such that

$$G_{I_1}(S(\mathbb{X}_n)) - G_{I_1}(S(\mathbb{X}_{n-1})) = \mathcal{R}_{I_1}(S(\mathbb{X}_{n-1}), x_n).$$

Thus it follows

$$\begin{aligned}
& \pi^{(I_1, 2, 1)}(S(\mathbb{X}_n)) - \pi^{(I_1, 2, 2)}(S(\mathbb{X}_n)) \\
&= \pi^{(I_1, 2, 1)}(S(\mathbb{X}_{n-1})) - \pi^{(I_1, 2, 2)}(S(\mathbb{X}_{n-1})) - (\pi^{(I_1)}(S(\mathbb{X}_{n-1})) + \pi^{(I_1)}(S(\mathbb{X}_{n-1, n})))\frac{x_n^2}{2} \\
&\quad + G(S(\mathbb{X}_n)) - G(S(\mathbb{X}_{n-1})).
\end{aligned}$$

where

$$G(S(\mathbb{X}_n)) = \sum_{I_1} C_{I_1} G_{I_1}(S(\mathbb{X}_n)).$$

Then following the notations

$$\begin{aligned}
\tilde{L}_1 &= \sum_{I_1} C_{I_1} \left(\pi^{(I_1, 2, 1)} - \pi^{(I_1, 2, 2)} \right) - G \\
f_n &= \tilde{L}(S(\mathbb{X}_n))
\end{aligned}$$

and it is obviously that $f(0) = 0$. Moreover since L_1 is additive, then $L_1(S(\mathbb{X}_{n-1})) + L_1(S(\mathbb{X}_{n-1,n})) = L_1(S(\mathbb{X}_n))$, and it follows

$$\begin{aligned} f_n &= f_{n-1} - \sum_{I_1} C_{I_1}(\pi^{(I_1)}(S(\mathbb{X}_{n-1})) + \pi^{(I_1)}(S(\mathbb{X}_{n-1,n}))) \frac{x_n^2}{2} \\ &= f_{n-1} - (L_1 S(\mathbb{X}_{n-1}) + L_1 S(\mathbb{X}_{n-1,n})) \frac{x_n^2}{2} \\ &= f_{n-1} - L_1 S(\mathbb{X}_n) \frac{x_n^2}{2}. \end{aligned}$$

By the telescoping sum of f_n , it holds that

$$f_n = \sum_{i=1}^n (f_i - f_{i-1}) + f_0 = \sum_{i=1}^n L_1 S(\mathbb{X}_i) \frac{x_i^2}{2}.$$

■

Theorem 4.1 (p-moment) *For any integer $p > 0$, there exist two linear functionals $L_p^{(1)} \in \mathcal{K}_1^{(p)}$, and $L_p^{(2)} \in \mathcal{K}_2^{(p)}$, such that for every path \mathbb{X} , the following equation follows:*

$$L_p^{(1)}(S(\mathbb{X})) = L_p^{(2)}(S(\mathbb{X})) = \sum_{i=1}^N x_i^p. \quad (7)$$

Obviously if (7) is true, then $L_p^{(1)}$ and $L_p^{(2)}$ are both additive.

Proof. Let's prove it by induction on p . It is true for $p = 1, 2$. Suppose that it holds for $p < P$. Let us study the case when $p = P$.

$$\begin{aligned} &\sum_{i=1}^N x_i^P \\ &= \sum_{i=1}^N \left(L_{p-2}^{(1)}(S(\mathbb{X})_i) - L_{p-2}^{(2)}(S(\mathbb{X})_{i-1}) \right) x_i^2 \\ &= \sum_{i=1}^N L_{p-2}^{(1)}(S(\mathbb{X})_i) x_i^2 - \sum_{i=1}^N L_{p-2}^{(2)}(S(\mathbb{X})_{i-1}) x_i^2 \end{aligned}$$

By Lemma 4.5, since $L_{p-2}^{(1)}$ is additive, then $\sum_{i=1}^N L_{p-2}^{(1)}(S(\mathbb{X})_{i-1}) x_i^2$ can be rewritten as a linear functional $G \in \mathcal{K}_p^{(1)}$ such that

$$G_1(S(\mathbb{X})_N) = \sum_{i=1}^N L_{p-2}^{(1)}(S(\mathbb{X})_i) x_i^2.$$

By Lemma 4.4, it follows that there exists $G_2 \in \mathcal{K}_p^{(2)}$, such that

$$G_2(S(\mathbb{X})_N) = \sum_{i=1}^N L_{p-2}^{(2)}(S(\mathbb{X})_{i-1}) x_i^2.$$

■

5 Multi-Dimensional Stream Case

The following lemma states that the empirical covariance of the increment of a multi-dimensional data stream can be fully characterized by its signatures.

Lemma 5.1 *Let $\mathbf{X} := \{X_n\}_{n=1}^L$ be a d -dimensional discretely sampled stream, $\{x_n\}_{n=1}^L$ be the associated increment process and \mathbb{X} be the corresponding lead-lag process of \mathbf{X} . For any $i_1, i_2 \in \{1, \dots, d\}$, there exists a linear functional L such that*

$$\sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)} = 2 \left(\pi^{(i_1, i_2+d)}(S(\mathbb{X})) - \pi^{(i_1, i_2)}(S(\mathbb{X})) \right).$$

Proof. For the case that $i_1 = i_2 = i$, it holds that

$$\sum_{n=1}^L (x_n^{(i)})^2 = \pi^{(i, i+d)}(S(\mathbb{X})) - \pi^{(i+d, i)}(S(\mathbb{X})) = 2(\pi^{(i, i+d)}(S(\mathbb{X})) - \pi^{(i, i)}(S(\mathbb{X}))).$$

as

$$\pi^{(i, i+d)} + \pi^{(i+d, i)} = \pi^{(i)} \pi^{(i+d)} = \pi^{(i)} \pi^{(i)} = \pi^{(i) \sqcup (i)} = 2\pi^{(i, i)}.$$

For the case that $i_1 \neq i_2$, the signature of the path $\mathbb{X}^{(i_1, i_2)}$, which is the (i_1, i_2) coordinate projection of \mathbb{X} is given as

$$S(\mathbb{X}^{(i_1, i_2)}) = \bigotimes_{n=1}^L \exp \left(x_n^{(i_1)} e_{i_1} + x_n^{(i_2)} e_{i_2} \right)$$

then it follows that

$$\pi^{(i_1, i_2)} S(\mathbb{X}) = \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} + \frac{1}{2} \sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)}.$$

the signature of the path $\mathbb{X}^{(i_1, i_2+d)}$, which is the (i_1, i_2+d) coordinate projection of \mathbb{X} is given as

$$S(\mathbb{X}^{(i_1, i_2+d)}) = \bigotimes_{n=1}^L \exp \left(x_n^{(i_1)} e_{i_1} \right) \otimes \exp \left(x_n^{(i_2+d)} e_{i_2+d} \right) \quad (8)$$

then it follows that

$$\pi^{(i_1, i_2+d)} S(\mathbb{X}) = \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} + \sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)}. \quad (9)$$

Combining (8) and (9), it follows that

$$\sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)} = 2(\pi^{(i_1, i_2+d)}(S(\mathbb{X})) - \pi^{(i_1, i_2)}(S(\mathbb{X}))).$$

■

Lemma 5.2 Let $\mathbf{X} := \{X_n\}_{n=1}^L$ be a d -dimensional discretely sampled stream, $\{x_n\}_{n=1}^L$ be the associated increment process and \mathbb{X} be the corresponding lead-lag process of \mathbf{X} . For any pairwise different $i_1, i_2, i_3 \in \{1, \dots, d\}$, there exists a linear functional L such that

$$\sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)} x_n^{(i_3)} = \frac{6}{5} \left(\pi^{(i_1, i_2, i_3)} + \pi^{(i_1, i_2, i_3+d)} - \pi^{(i_1+d, i_2, i_3+d)} - \pi^{(i_1, i_2+d, i_3+d)} \right) (S(\mathbb{X})).$$

Proof. The signature of the path $\mathbb{X}^{(i_1, i_2, i_3)}$, which is the (i_1, i_2, i_3) coordinate projection of \mathbb{X} is given as

$$S(\mathbb{X}^{(i_1, i_2, i_3)}) = \bigotimes_{n=1}^L \exp \left(x_n^{(i_1)} e_{i_1} + x_n^{(i_2)} e_{i_2} + x_n^{(i_3)} e_{i_3} \right)$$

then it follows that

$$\begin{aligned} \pi^{(i_1, i_2, i_3)} S(\mathbb{X}) &= \sum_{n_1 < n_2 < n_3} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_3}^{(i_3)} + \frac{1}{2} \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_2}^{(i_3)} \\ &\quad + \frac{1}{2} \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_2}^{(i_3)} + \frac{1}{6} \sum_{n_1=1}^L x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_1}^{(i_3)}. \end{aligned}$$

The signature of the path $\mathbb{X}^{(i_1, i_2+d, i_3+d)}$, which is the (i_1, i_2, i_3) coordinate projection of \mathbb{X} is given as

$$S(\mathbb{X}^{(i_1, i_2+d, i_3+d)}) = \bigotimes_{n=1}^L \left(\exp \left(x_n^{(i_1)} e_{i_1} \right) \otimes \exp \left(x_n^{(i_2)} e_{i_2+d} + x_n^{(i_3)} e_{i_3+d} \right) \right).$$

then it follows that

$$\begin{aligned} \pi^{(i_1, i_2+d, i_3+d)} S(\mathbb{X}) &= \sum_{n_1 < n_2 < n_3} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_3}^{(i_3)} + \frac{1}{2} \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_2}^{(i_3)} \\ &\quad + \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_2}^{(i_3)} + \frac{1}{2} \sum_{n_1=1}^L x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_1}^{(i_3)}. \end{aligned}$$

Similarly we have that the signature of the path $\mathbb{X}^{(i_1, i_2, i_3+d)}$, which is the $(i_1, i_2, i_3 + d)$ coordinate projection of \mathbb{X} is given as

$$S(\mathbb{X}^{(i_1, i_2, i_3+d)}) = \bigotimes_{n=1}^L \left(\exp \left(x_n^{(i_1)} e_{i_1} + x_n^{(i_2)} e_{i_2} \right) \otimes \exp \left(x_n^{(i_3)} e_{i_3+d} \right) \right).$$

and thus it holds that

$$\begin{aligned} \pi^{(i_1, i_2, i_3+d)} (S(\mathbb{X})) &= \sum_{n_1 < n_2 < n_3} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_3}^{(i_3)} + \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_2}^{(i_3)} \\ &\quad + \frac{1}{2} \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_2}^{(i_3)} + \frac{1}{2} \sum_{n_1=1}^L x_{n_1}^{(i_1)} x_{n_1}^{(i_2)} x_{n_1}^{(i_3)}. \end{aligned}$$

Moreover we have that

$$\pi^{(i_1+d, i_2, i_3+d)}(S(\mathbb{X})) = \sum_{n_1 < n_2 < n_3} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_3}^{(i_3)} + \sum_{n_1 < n_2} x_{n_1}^{(i_1)} x_{n_2}^{(i_2)} x_{n_2}^{(i_3)}.$$

Combining the above equations, it follows that

$$\sum_{n=1}^L x_n^{(i_1)} x_n^{(i_2)} x_n^{(i_3)} = \frac{6}{5} \left(\pi^{(i_1, i_2, i_3)} + \pi^{(i_1, i_2, i_3+d)} - \pi^{(i_1+d, i_2, i_3+d)} - \pi^{(i_1, i_2+d, i_3+d)} \right) (S(\mathbb{X})).$$

■

6 Numerical Examples

6.1 Toy Example 1: Correlation estimation

In this toy example, we want to demonstrate that the signature of a stream can be used as a basis function to represent standard statistics, for example, the mean and the covariance matrix of the increment process.

Example 6.1 We simulate 400 samples of the pair $\{\rho_n, \mathbb{X}_{\rho_n}\}_{n=1}^{N=400}$, where ρ_n is iid and uniformly distributed in $[0, 1]$, and for each ρ_n , \mathbb{X}_{ρ_n} is generated as a 2-dimensional random walk of length L with the correlation ρ_n , i.e.

$$x_{\rho_n} \stackrel{iid}{=} \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix}\right).$$

How can we estimate the model parameter ρ for each sample path?

Our method is simply to do the linear regression of the correlation parameter against the truncated signature of the sample path. To better judge the performance of our method, we used the empirical correlation as a benchmark. The empirical correlation for each sample path \mathbb{X}_ρ is defined as follows:

$$\hat{\rho} = \frac{\sum_{n=0}^{L-1} \left(x_\rho^{(1)}(n) - \bar{x}_\rho^{(1)} \right) \left(x_\rho^{(2)}(n) - \bar{x}_\rho^{(2)} \right)}{\sqrt{\sum_{n=0}^{L-1} \left(x_\rho^{(1)}(n) - \bar{x}_\rho^{(1)} \right)^2 \sum_{n=0}^{L-1} \left(x_\rho^{(2)}(n) - \bar{x}_\rho^{(2)} \right)^2}}$$

Some parameters I chose are given as follows:

$$L = 120, N = 200, d = 3$$

Figure 2 shows that the empirical correlation is better in terms of MSE, especially when ρ is near $+1$ and -1 . However due to the nature of polynomial regression, the signature-approach perform worse when ρ is near the boundary. However the reason why the signature approach is not satisfactory is not because that the truncated signature do not include enough information of the path. Instead the reason is that the regression method we used is too simple and it should be combined with advanced non-linear regression techniques, e.g. rational regression or some local regression methods. Theoretically if properly combined with advanced regression

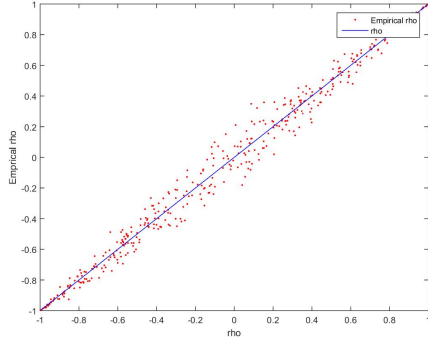


Figure 1: The plot of the empirical correlation v.s the actual correlation

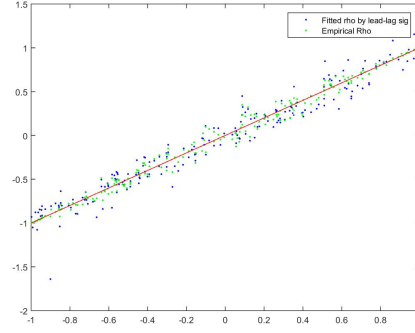


Figure 2: The plot of two estimated correlation against the actual correlation.

techniques, we should be able to recover the empirical correlation. It is because that by definition of the signature of a stream, Lemma 5.1 shows that the empirical covariance/variance of the increment process is a linear combination of the truncated signature up to degree 2, and the ratio of the empirical covariance and the square root of empirical variance of two coordinate increments gives the empirical correlation.

6.2 Toy Example 2: Using signatures to classify two classes of random walks

Example 6.2 Let \mathbb{X} denote a standard 3-dimensional random walk of length L , and \mathbb{Y} denote the other random walk, where $y^{(1)}, y^{(2)}$ are independent and move to $+1$ and -1 with probability 0.5, but $y^{(3)} = y^{(1)}y^{(2)}$. Given one realization of a random walk of length L generated either by the distribution of \mathbb{X} or that of \mathbb{Y} , which distribution this realized path is from?

In this example, we can't distinguish which distribution one sample path is generated from by looking at its empirical mean and covariance matrix of the increment distribution, it is simply because that

$$\begin{aligned}\mathbb{E}[x] &= \mathbb{E}[y] = 0; \\ \text{cov}[x] &= \text{cov}[y] = I_3.\end{aligned}$$

But we can almost perfectly classify this sample path using the truncated signatures in this case. We summarize the procedure as follows:

1. We simulate N paths based on the distribution of \mathbb{X} and \mathbb{Y} respectively.
2. Compute the truncated signature of those sample paths up to degree d .
3. For each sample path \mathcal{X} , let the response variable define in the following way:

$$f(\mathcal{X}) = \begin{cases} 1 & \text{if } \mathcal{X} \text{ is sampled from } \mathbb{X}; \\ 0 & \text{if } \mathcal{X} \text{ is sampled from } \mathbb{Y}. \end{cases}$$

4. We randomly select half of the dataset as the learning set, and the rest data as the backtesting set. Apply SVM classification method to $f(\mathcal{X})$ against $S(\mathcal{X})_d$ in the learning set, where $d = 3$.
5. After obtaining the classifier \hat{f} , for any new given path \mathcal{X}^* , by plugging it to the classifier \hat{f} , the estimated class of \mathcal{X}^* is given by $\hat{f}(\mathcal{X}^*)$.

In this example, we choose $N = 200$, $L = 100$ and $d = 3$. The incorrect selection ratio is $1/400$, and it means that there is only one mis-classification for the whole dataset of size 400. It is noted that the sample space of \mathbb{Y} is actually the subspace of the sample space of \mathbb{X} , and theoretically if \mathcal{X} is in the sample space of \mathbb{X} , its category is not distinguishable from this sample path trajectory.

7 Appendix

References

- [1] Guy Flint, Ben Hambly, and Terry Lyons. Discretely sampled signals and the rough hof process. *arXiv preprint arXiv:1310.4054*, 2013.
- [2] Terry Lyons, Thierry Lévy, and Michael Caruana. *Differential Equation driven by Rough Paths*. Springer, 2006.